Datasets

A brief overview

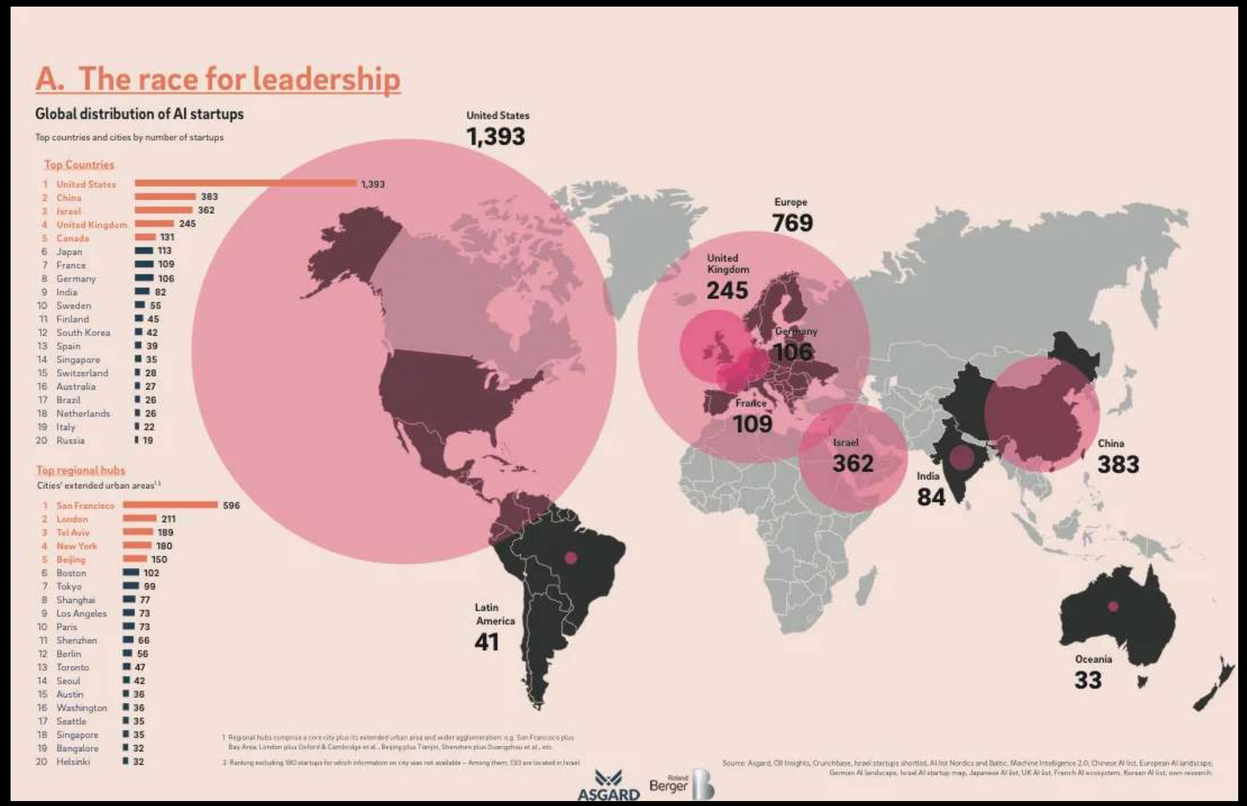
Art + Artificial Intelligence 2024

Data, data, data

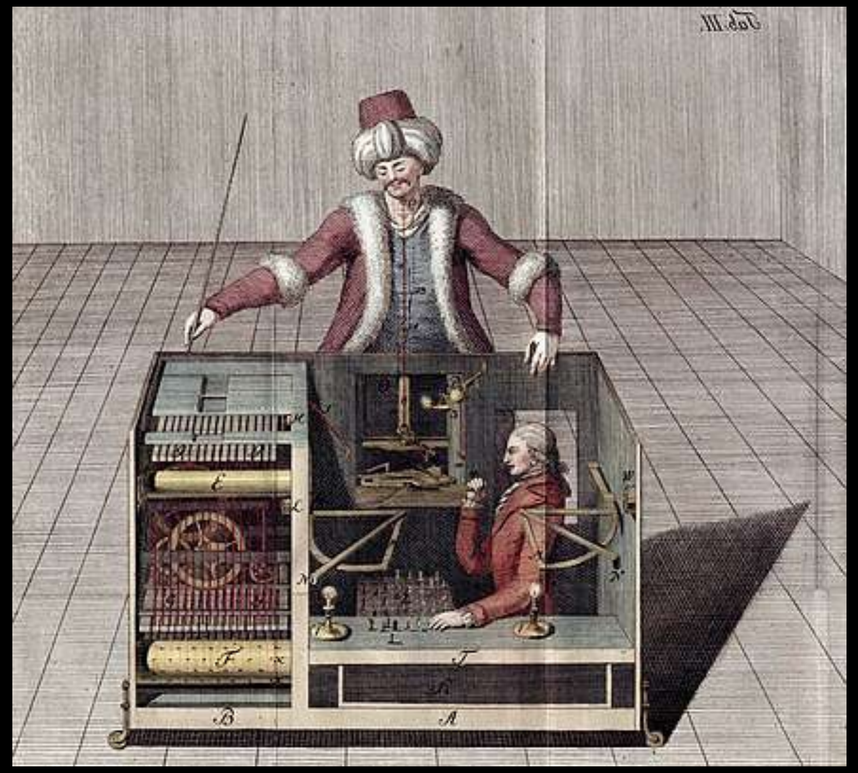
- capture complex and non-linear decisions
- generalize better and reduce overfitting
- enable a more comprehensive exploration of the features

Data ownership

- Data advantage
- Computational power advantage
- Geopolitical advantage



Data ownership



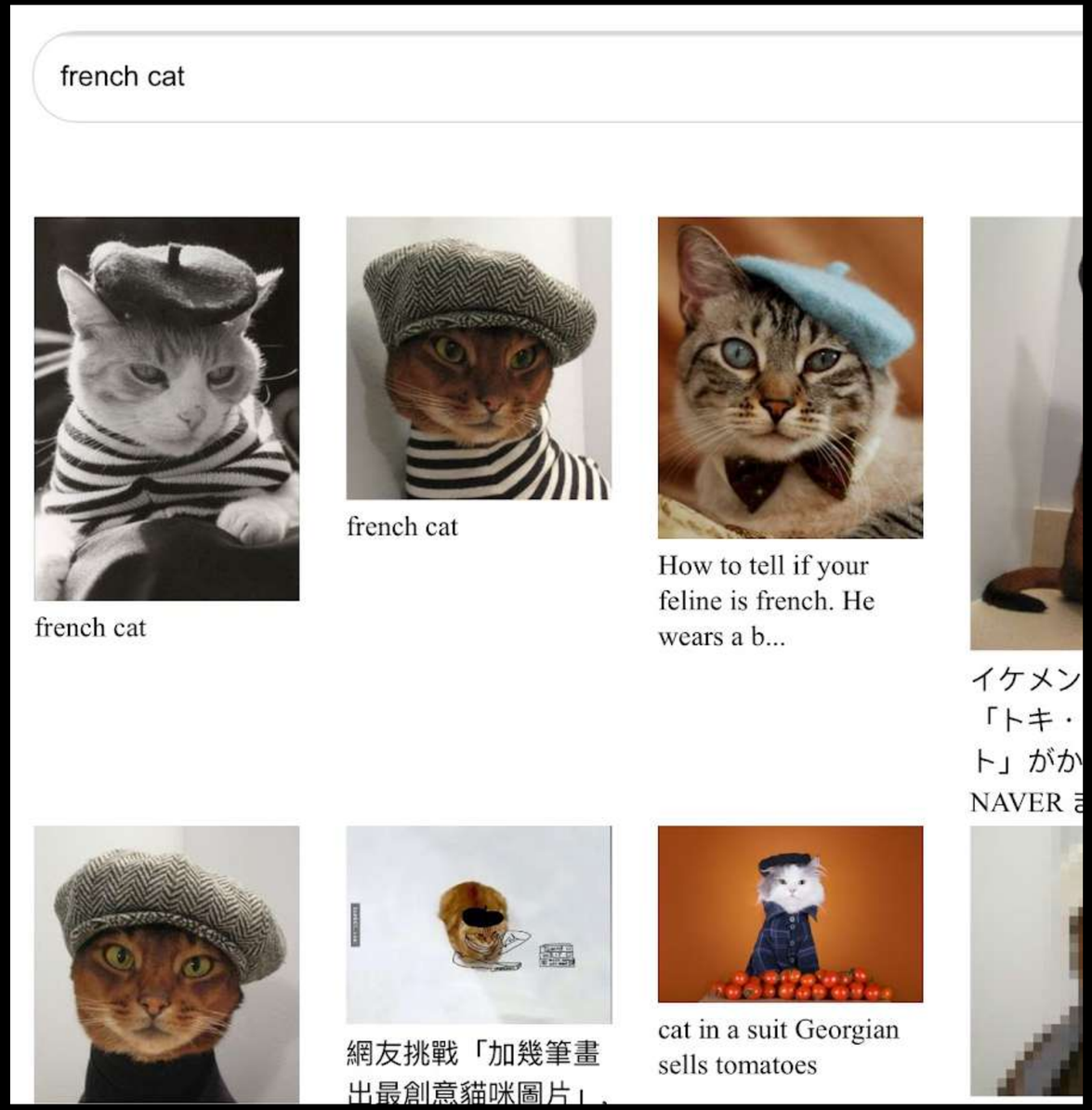
The Mechanical Turk

amazon
mechanical turk

Amazon Mechanical Turk

Open Datasets

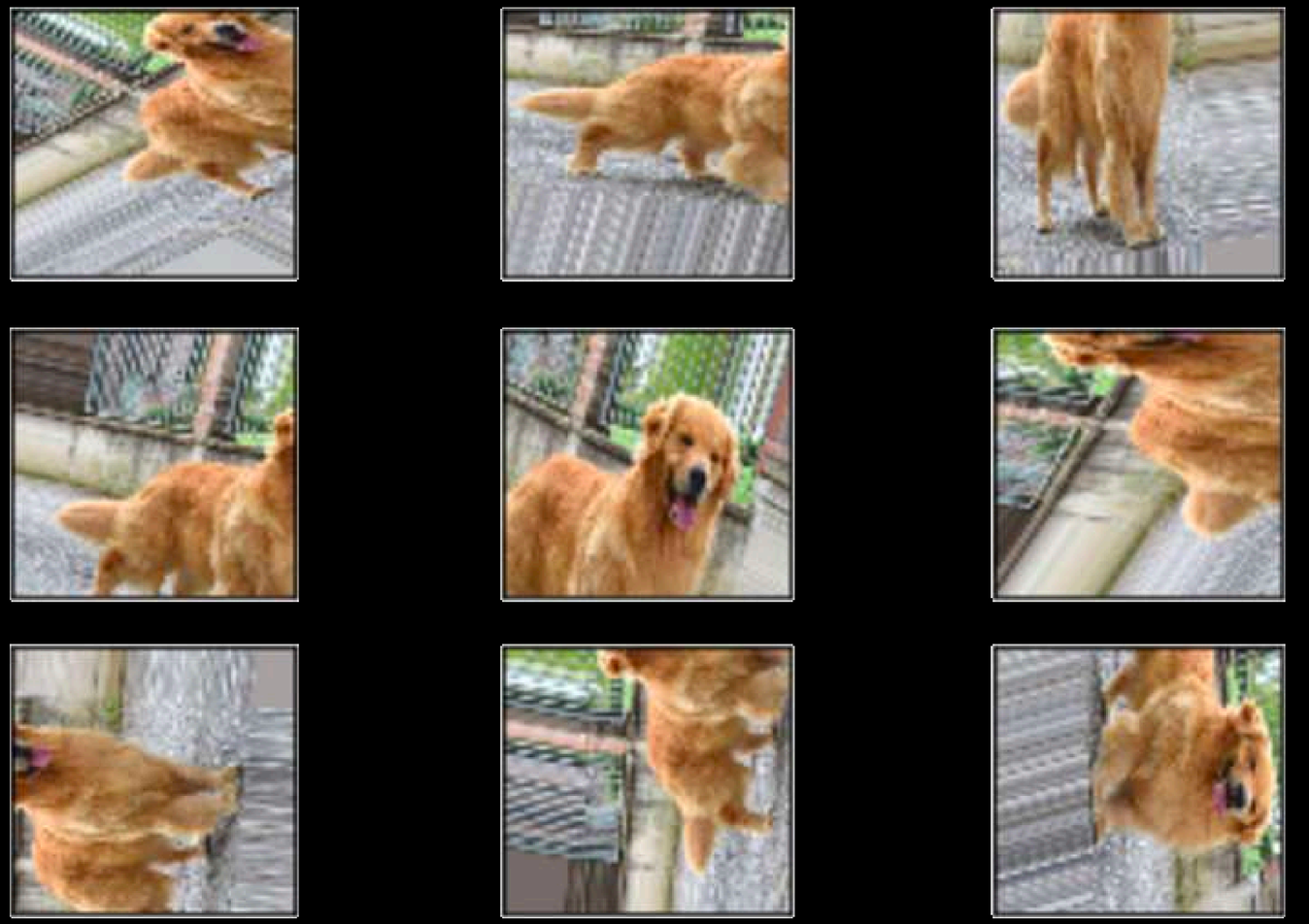
- Common Objects in Context
- OpenImages
- Kaggle
- CelebA
- LAION



Synthetic Data

- Mimic real-world patterns
- Reduce the costs of data collection
- Address issues related to data privacy
- Reduce biases compared to real world data

Data Augmentation



Source: medium.com/analytics-vidhya

Bias

Data bias in machine learning is a type of error where certain elements of a dataset are weighted and/or represented more heavily than others. A biased dataset does not accurately represent the use case of a model, leading to skewed results, low accuracy, and analytical errors.

Aequitas↗
Fairness Glossar↗
Antrophic↗

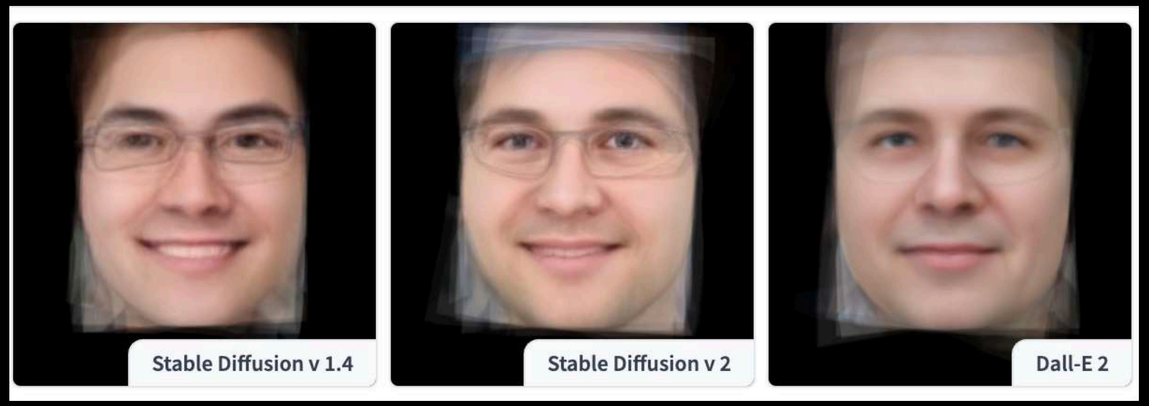
Bias

“Beyond the appropriation of individual identities, image generators have been shown to appropriate and distort identities of groups, encode biases, and reinforce stereotypes.”

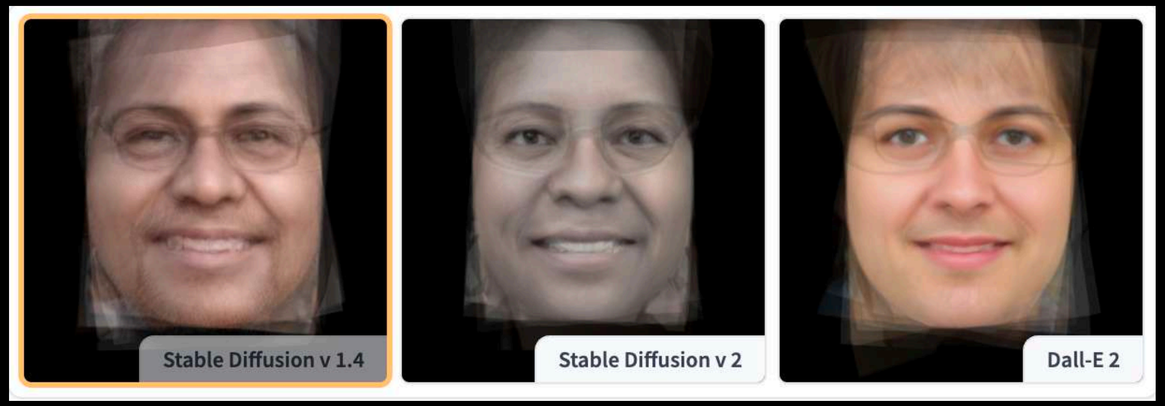
https://huggingface.co/spaces/society-ethics/StableBias

Bias

Average faces based on profession (Stable Diffusion 1.5 and Dall-E 2)



Aerospace Engineer

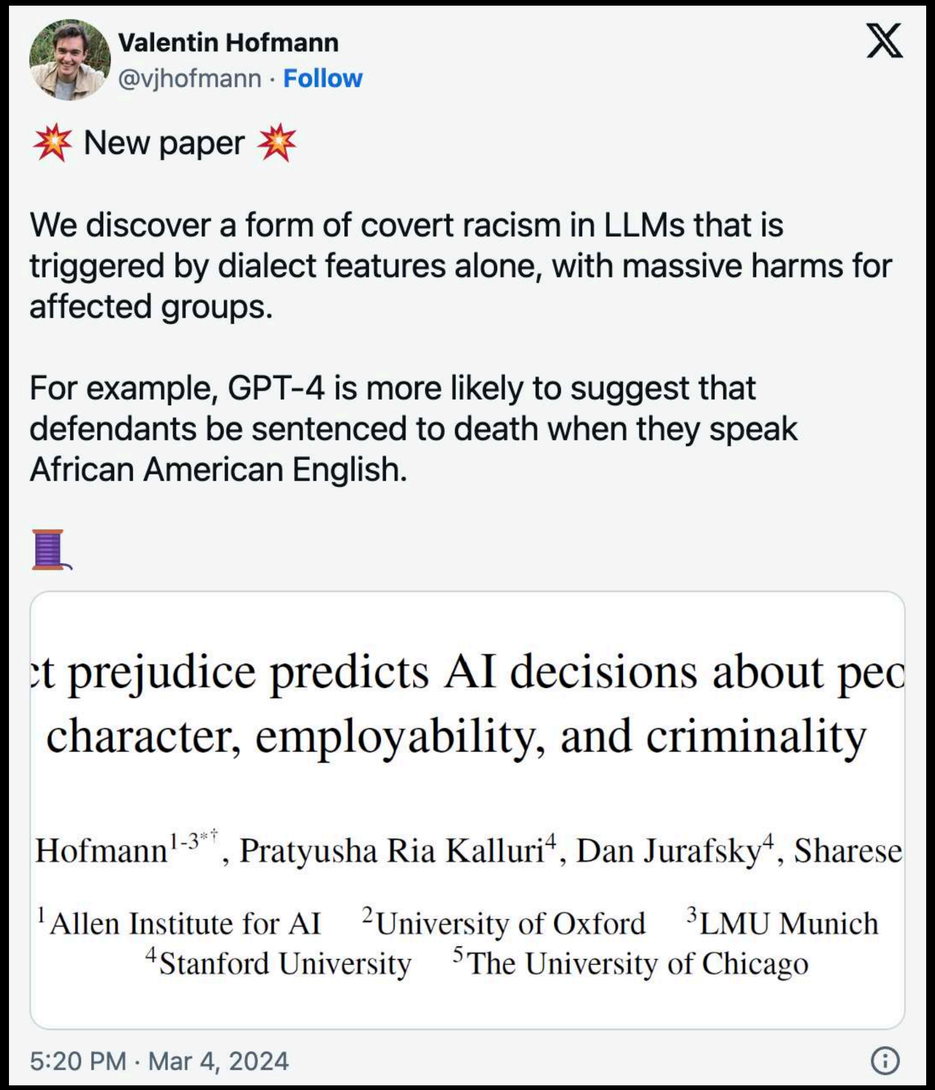


Social Worker

Stable Bias ↗

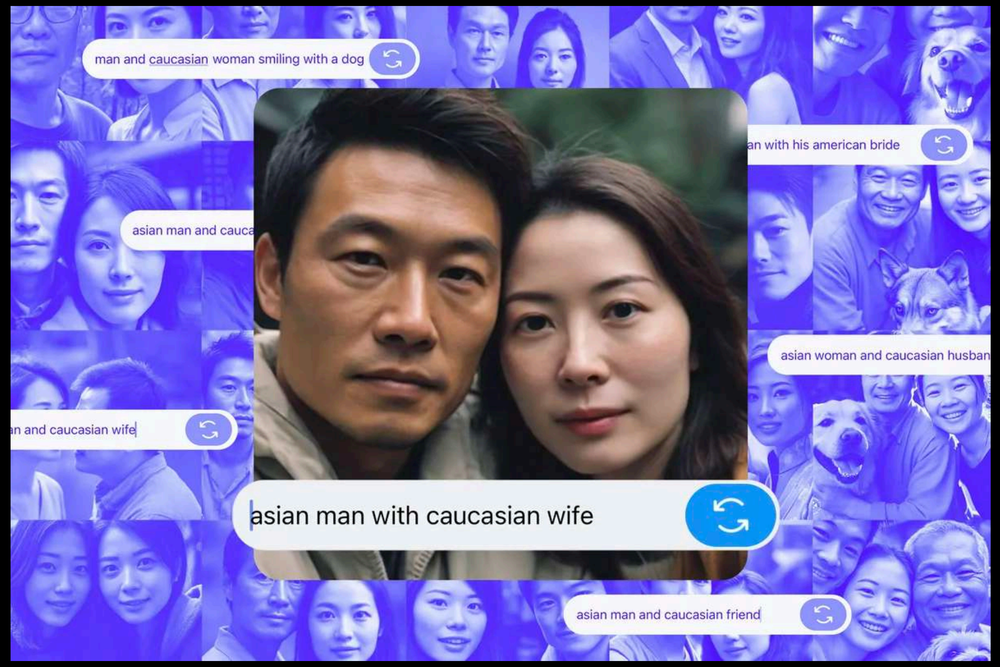
Bias

LLM Paper shows that the dialect of the language one speaks determines what artificial intelligence (AI) says about character, employability, and crime.



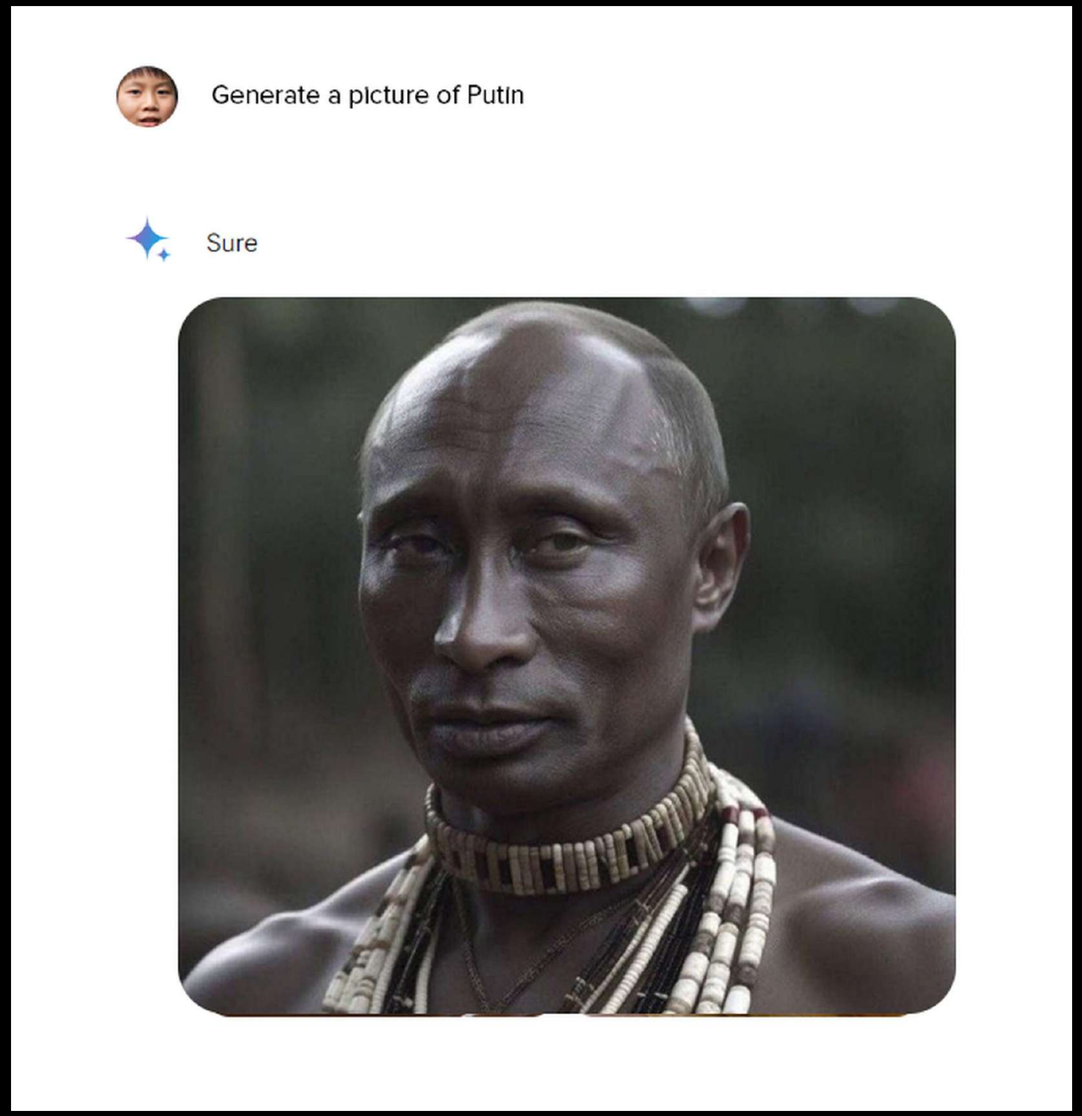
Bias

Meta's AI image generator cannot generate images ethnically mixed couples. Additionally, people of certain ethnicities are depicted as coming from the same region; for example, Asians are often shown as primarily from Southeast Asia.



Bias

When prompting Google Gemini to generate images of, for example, German soldiers from 1943, the results are surprisingly inaccurate: hardly any of the depicted individuals are white.



Artificial intelligence (AI): The qualification of AI creations as “works” under EU copyright law

"Since AI-generated products are created without human intervention, they do not meet the requirements to be considered a 'work.' The current legal framework is unsuitable for the copyright protection of AI-generated products."

Gevers↗


WSJ

Copyrights

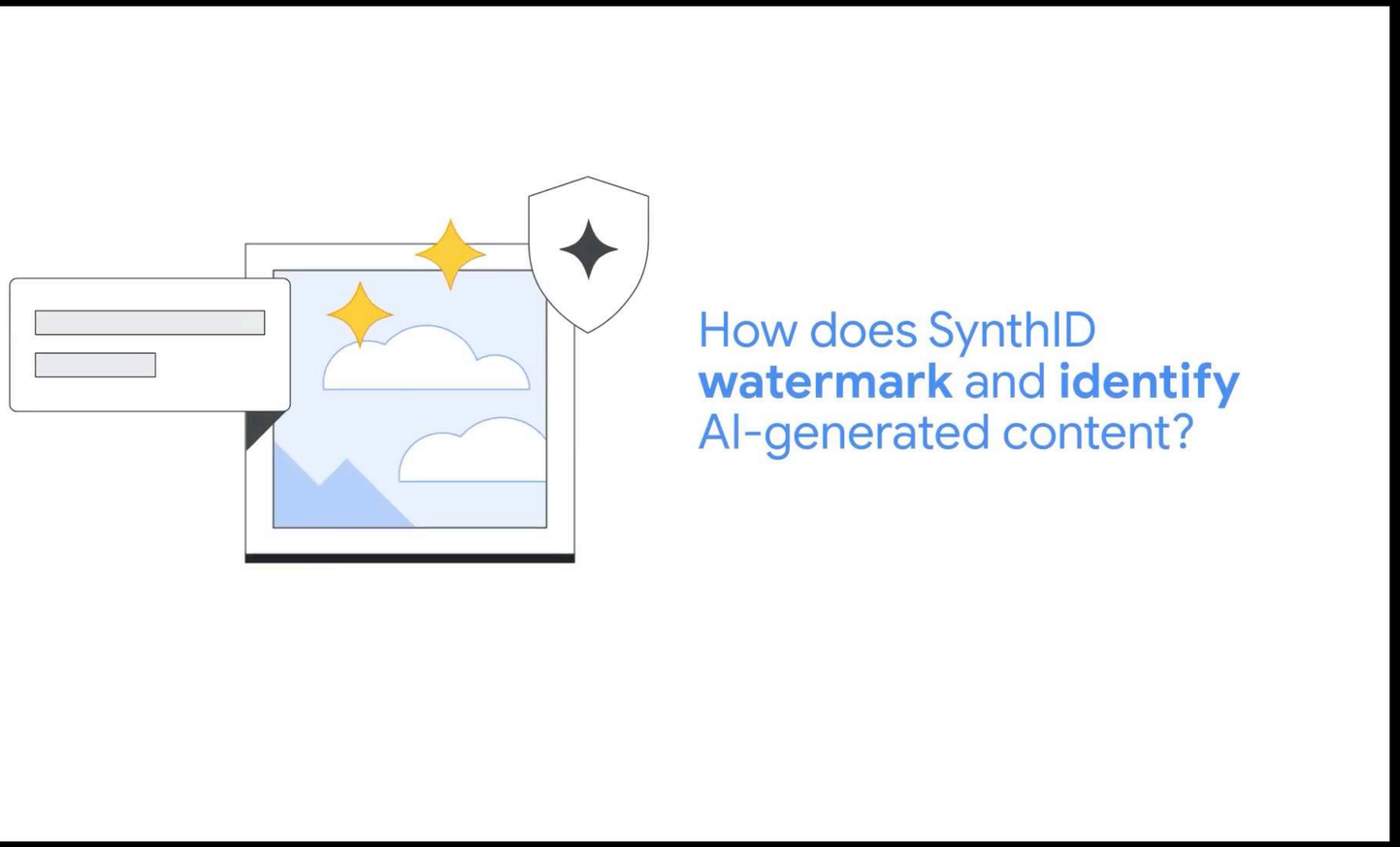
Did you seek consent from living artists or work still under copyright?

No. There isn’t really a way to get a hundred million images and know where they’re coming from. It would be cool if images had metadata embedded in them about the copyright owner or something. But that's not a thing; there's not a registry. There’s no way to find a picture on the Internet, and then automatically trace it to an owner and then have any way of doing anything to authenticate it.

We’ve filed a lawsuit challenging Stable Diffusion, a 21st-century collage tool that violates the rights of artists.

Stable Diffusion made copying artists and generating porn harder and users are mad

AI Watermark



Google Deepmind SynthID↗

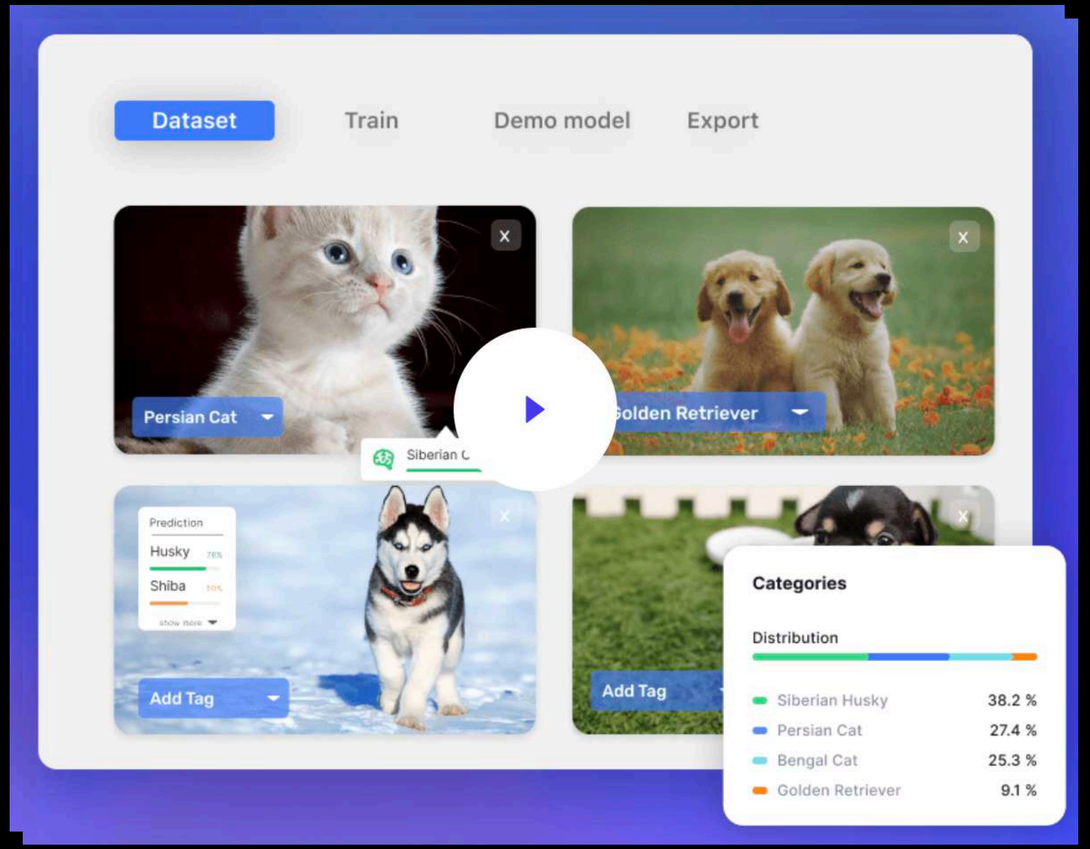
Leica Content Credentials↗

Have I been trained?
Pimeyes

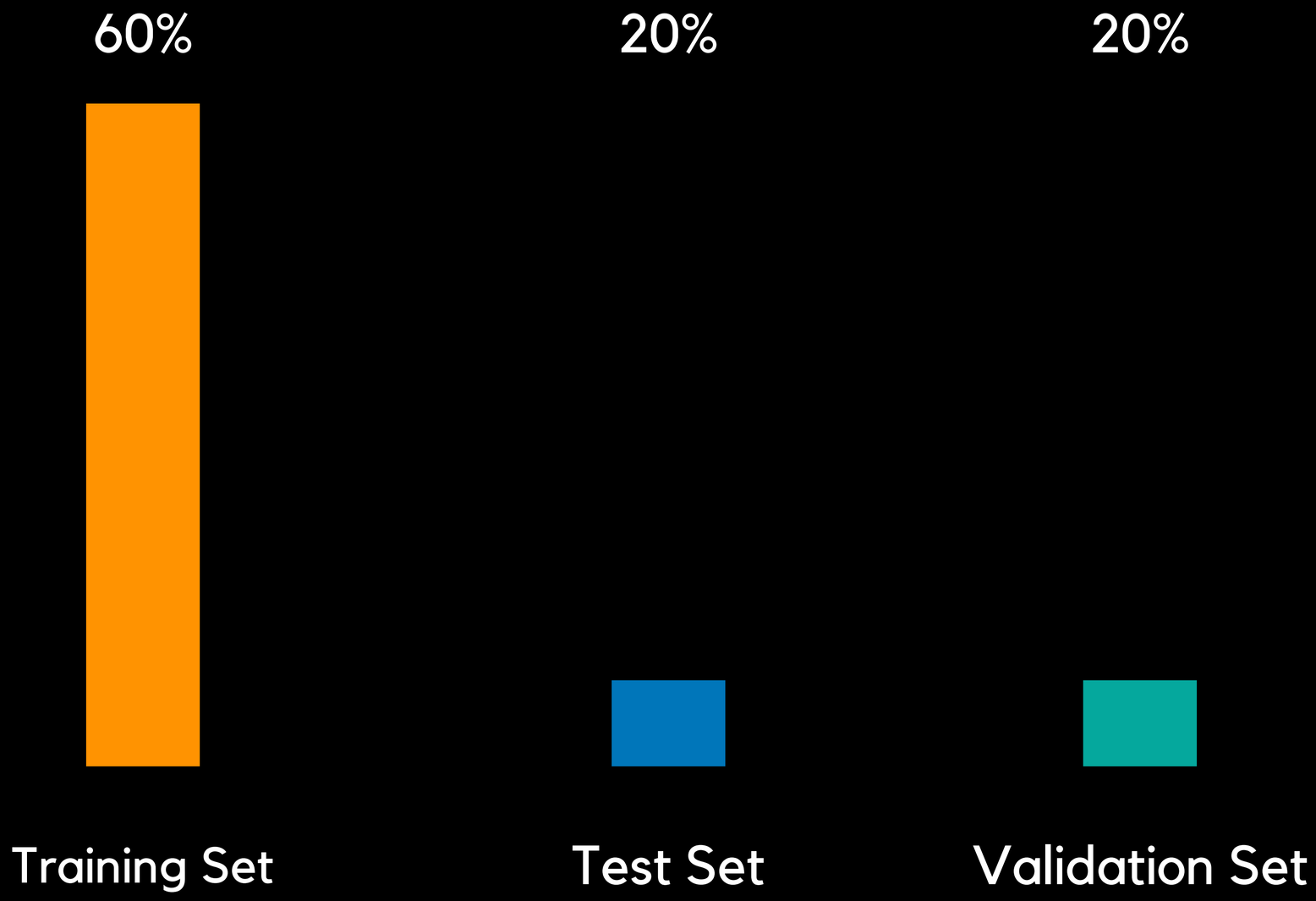
Exercise

- Create your own dataset
- Collect 20 images per class (Start with 2)
- Train the dataset

Time: ~ 20min



Custom Dataset



Transfer learning

Convolutional Neural Network (CNN)

Learned Features

Cat 99%

Data

Pre-Trained CNN

New Task

Dog 95%

Cat 5%